图书情报工作
第 64 卷 第 12 期 2020 年 6 月

文本相似度视角下我国大数据政策比较研究*

■ 张涛[1] 马海群[2] 易扬[3]
[1] 黑龙江大学信息与网络中心 哈尔滨 150080 [2] 黑龙江大学信息资源管理研究中心 哈尔滨 150080
[3] 黑龙江大学数学科学学院 哈尔滨 150080

摘 要：[目的/意义]大数据政策的制定与实施是国家推动大数据产业发展的重要手段，因此对大数据的政策研究也受到了社会广泛关注。[方法/过程]以文本相似度为视角对国务院发布的《促进大数据发展行动纲要》和我国 22 个地区发布的大数据政策文本进行比较研究。[结果/结论]数据表明：广东省、福建省所制定的政策最为完整和全面，数据开放共享和安全保障在各地区大数据政策制定层面整体关注最高，呈现出相似性，在内蒙古自治区、四川省等地区大数据政策制定中区域特色较为突出，呈现出差异性。随着各地区相继颁布人工智能政策，未来对人工智能视域下大数据政策的研究将成为新方向。

关键词：文本相似度 大数据政策 政策比较研究 政策文本计算
分类号：D63 TP309.2
DOI：10.13266/j.issn.0252-3116.2020.12.004

21 世纪初，在 Michchael Laver 等提出政策文本计算的基本概念后，大量的计算机科学理论与方法开始被运用于海量文本挖掘和文本计算分析，政策文本计算的出发点是对政策文本的自然语言处理[1]，而文本相似度计算是政策文本计算中重要研究方法之一。随着技术的进步，文本相似度计算的精确度也在不断提升，并被广泛应用于文献查重、智能机器问答、文本智能分类等领域，对文献调研发现，目前少有研究者将其用于政策比较研究中。自 2015 年 8 月 31 日国务院发布《促进大数据发展行动纲要》（以下简称《纲要》）以来，国家各部委、各地区基于《纲要》内容先后出台了一系列政策来推动大数据产业的发展。虽然大数据政策制定的总体目标相同，但由于区域特色不同，各地区所出台的政策差异较大，因此对国家与各地区间大数据政策的比较研究显得尤为重要，研究不但可以通过政策间的相似性探寻重点关注内容，还可以通过政策间的差异性探寻区域发展特色。目前定性政策研究方法存在效率较低及主观性较强的问题，因此，本文以文本相似度为视角对我国大数据政策进行比较研究，以实现对不同地区间政策文本科学化分析，进而实现为政府决策提供支持的目标。

1 文献综述

目前，国内外学界形成了一系列大数据政策比较研究的成果，在国内，2014 年，张勇进等[2]通过对国外政府大数据政策的调研，从 3 个层面比较分析发达国家大数据政策，并总结了其共性特点，最终形成大数据政策比较研究框架。2017 年，汤志伟等[3]基于工具维度和评价维度对中美开放政府数据政策进行比较研究。2017 年，王本刚和马海群[4]对西方发达国家的开放数据政策进行比较研究，基于对国外政策的研究，提出我国政府开放数据政策应坚持的原则和需要采取的措施。2019 年，赵远[5]利用内容分析法和比较研究法，从政策工具和政策目标两个维度对我国大数据发展指数靠前的省市进行比较分析。在国外，2014 年，A. Zuiderwijk 等[6]提出开放数据政策框架应包括环境因素、政策内容、绩效指标和公共价值，并以荷兰政府为例比较了开放政府数据政策之间的相似和差异之处。2016 年，E. Chatzinikolaou 等[7]对希腊生命观察研究基础设施（简称 LWG RI）数据政策背后的基本原

* 本文系黑龙江省哲学社会科学研究规划项目"智能+视阈下基于语料库的数据政策模型构建与实证研究"（项目编号：19TBQ073）和国家社会科学基金重点项目"开放数据与数据安全的政策协同研究"（项目编号：15ATQ008）研究成果之一。
作者简介：张涛（ORCID：0000-0002-3367-4541）高级工程师，博士研究生，E-mail：zhangtao@hlju.edu.cn；马海群（ORCID：0000-0002-2091-7620）教授，博士，博士生导师；易扬（ORCID：0000-0001-5496-644X），硕士研究生。
收稿日期：2019-09-30 修回日期：2020-01-31 本文起止页码：26-37 本文责任编辑：王传清

理、共享研究数据当前的法律情况、数据所有者/提供商与 LWG RI 签署的数据共享协议等方面进行了详细描述,并进行综合分析。2019 年,D. Tatiana-Camelia[8] 提出基于实证的政策制定,证实了该方法在政策比较研究中起到了十分重要的作用。

综上所述,在国内外对数据政策的研究成果中,尚无学者从文本相似度视角对大数据政策进行比较研究。因此,本文通过计算《纲要》与地方省级政府发布的大数据政策文本相似度的数值,对国家与各地区大数据政策的相似性和差异性进行比较分析,并最终提出我国大数据产业发展的政策建议。

2 研究方法

文本相似度计算在不同领域中发挥着重要作用并被广泛应用,由于其应用场景不同,内涵和计算方法也有所差异。D. Lin[9] 从信息论的角度阐明文本相似度的共性和差异关系,共性越大、差异越小,则相似度越高;共性越小、差异越大,则相似度越低;通过以上理论可以假定政策文本间相似度值越高,则共性越大。基于陈二静等[10]、黄文彬等[11]、李琳等[12]对文本相似度计算方法的综述可知,文本相似度计算较为常用的是词袋模型和词向量模型,词袋模型用于规范性文本、短文本效果较好,而词向量模型适用于对大规模文本的分析。经过分析,本文研究对象具有以下特点:①政策文本具有语言精炼、规范、严谨等特点。②研究的政策文本数据集较小。③研究对象都属于大数据政策,其中特征词较为一致。④计算对象为短文本。基于以上因素,在借鉴曹祺等[13]对 TF-IDF 模型、LSA 模型、LDA 模型和 Doc2Vec 模型的对比分析的基础上,选取 Doc2Bow 与 TF-IDF 相结合方法对政策文本相似度进行计算。

2.1 文本组织结构

政策文本从组织结构上是由文本、语句、词语构成,如图 1 所示:

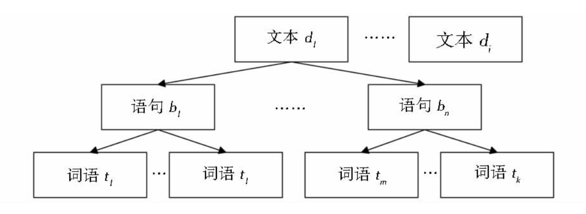


图 1 文本组织结构的树状层次

文本集合 D 为:$D=\{d_1,d_2,\cdots,d_i\}$,i 为文本的数量。

设 d_i 为文本集 D 的一个文本,b_i 为文本 d_i 中的语句,t_i 为语句 b_i 中的词语,文本 d_i 和语句 b_i 可用如下形式描述:

$$d_i=\{d_i,0,d_i,1,\cdots,d_i,j,\cdots,d_i,n-1\} \tag{1}$$

$$d_i=(b_1,b_2,\cdots,b_i,\cdots,b_n) \tag{2}$$

$$b_i=(t_1,t_2,\cdots,t_i,\cdots,t_k) \tag{3}$$

其中 n 表示文本 d_i 中语句的数量,k 表示语句 b_i 中词语的数量。

2.2 文本相似度计算过程

假设任意两个文本 d_1、d_2,分别表示为公式(4)和公式(5):

$$d_1=(b_{11},b_{12},\cdots,b_{1i},\cdots,b_{1m}) \tag{4}$$

$$d_2=(b_{21},b_{22},\cdots,b_{2i},\cdots,b_{2n}) \tag{5}$$

m 和 n 分别为文本中语句的数量。将文本中的语句视为短文本,重点对语句间的相似度进行计算。设 d_{12} 为 d_1、d_2 的相似度矩阵,则 d_{12} 可表示如下:

$$d_{12}=d_1^{\,T}\times d_2=\begin{Bmatrix} b_{11}b_{21} & b_{11}b_{22} & \cdots & b_{11}b_{2n} \\ b_{12}b_{21} & b_{12}b_{22} & \cdots & b_{12}b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{1m}b_{21} & b_{1m}b_{22} & \cdots & b_{1m}b_{2n} \end{Bmatrix} \tag{6}$$

取公式(6)中任意一项,如 $b_{11}b_{12}$ 来分析,具体实现步骤如下:

第一步:对所要分析的政策文本按语句拆分并分词,见公式(7)和公式(8)。

第二步:将分词后的短文本利用 doc2bow 方法转换为稀疏向量。

第三步:利用 TF-IDF 模型将政策文本进行处理得到 TF-IDF 值。

第四步:通过所计算 TF-IDF 值,利用余弦相似度计算语句相似性。

$$b_{1m}=(t_{11},t_{12},\cdots,t_{1i},\cdots,t_{1m}) \tag{7}$$

$$b_{2n}=(p_{11},p_{12},\cdots,p_{1i},\cdots,p_{1n}) \tag{8}$$

其中 t_{1m}、p_{1n} 分别是上述语句词向量,用其夹角余弦值来表示距离,计算两个向量的余弦值来表示两个语句相似度距离[14],取值范围从 0 到 1 之间,数值越大,则相似度越高。从而得出其语句相似度数值,见公式(9)。

$$similarity=\cos(b_{1m},b_{2n})=\frac{b_{1m}\cdot b_{2n}}{\|b_{1m}\|\,\|b_{2n}\|}=\frac{\sum_{m=1}^{m}\sum_{n=1}^{n}t_{1m}p_{1n}}{\sqrt{\sum_{m=1}^{m}t_{1m}^{2}}\sqrt{\sum_{n=1}^{n}p_{1n}^{2}}} \tag{9}$$

3 研究过程

本文的主要研究对象有两类:一是《纲要》,该文件是国务院发布大数据产业布局的战略性政策,是目前促进大数据发展第一份权威性、系统性文件,《纲要》中提到了三大主要任务和十项工程,三大主要任务是大数据政策执行的核心部分,也是各地区制定大数据政策的重要参考,主要包括:①任务 1:加快政府数据开放共享,推动资源整合,提升治理能力(以下简称数据开放);②任务 2:推动产业创新发展,培育新兴业态,助力经济转型(以下简称创新发展);③任务 3:强化安全保障,提高管理水平,促进健康发展(以下简称安全保障)。因此,本文将此部分内容提取并形成框架,把其作为比较对象。二是 2013 - 2019 年我国 22 个地区发布的省级政府层面大数据行动计划或实施方案,将这些政策文件作为被比较研究对象。现对整个研究过程分解如图 2 所示:

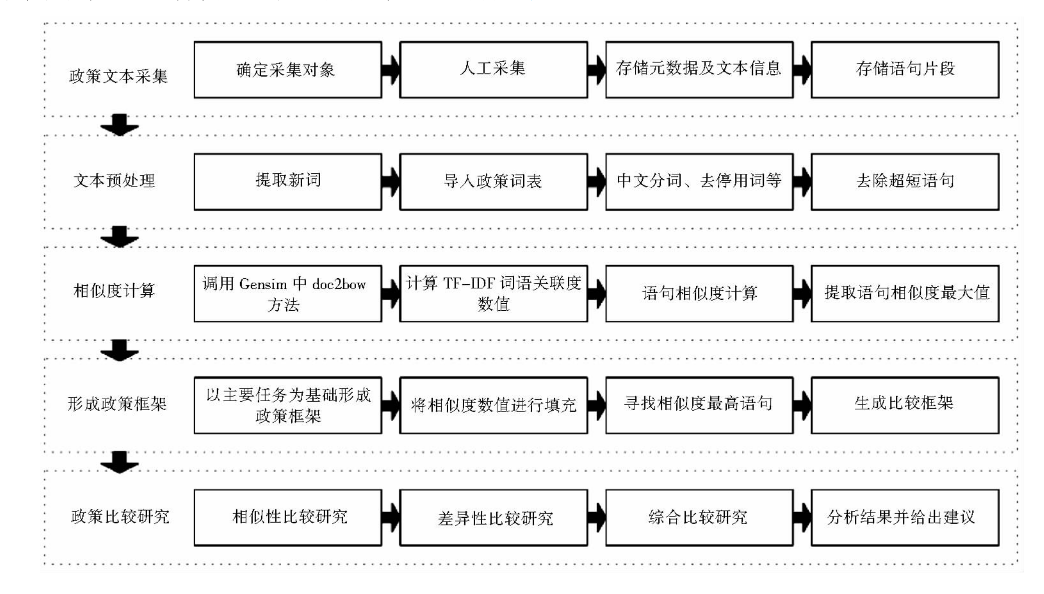


图 2 基于文本相似度的政策比较研究过程

3.1 政策文本采集

确定采集对象是政策研究的初始环节,把所研究的政策通过人工采集的方式录入到自建语料库[15]中来,将政策文本信息分为三部分进行存储:①元信息:主要记录发布时间、发布机构、有效时间、政策类别等字段;②内容信息:以 *. txt 文本文档(UTF - 8 格式)的形式存储在服务器中;③语句片段信息:数据库中会按句子划分进行语句单元存储。

3.2 文本预处理

利用中国科学院 ICTCLAS[16] 中的新词提取功能对政策文本进行关键词获取,结合自建语料库中原有政策词表,形成 334 条政策词语并导入,通过 Python 语言中的 jieba 工具对文本进行分词、去停用词等预处理操作,将文本数据转换为可分析处理的初始格式。由于超短语句(分词后字符小于 5 的语句)对所计算的结果影响较大,因此需要去除此类无效语句。文本预处理是相似度计算最重要的环节之一,最终计算结果的精准度与该过程密切相关[17]。

3.3 相似度计算

本文将政策文本按语句片段划分为短文本,共分为 5 678 条语句,总计 338 122 字符数,将其作为比较研究对象,利用 Python 语言 Gensim 工具中的 BOW 模型和 TF-IDF 模型[18],按照 2.2 小节中的文本相似度计算方法,对《纲要》中三大任务与各地区大数据政策文本进行语句相似度计算。从语句层面看,相似度数值越高,则政策间的共性越大。

3.4 生成政策框架

抽取《纲要》中三大任务形成大数据产业发展任务一级指标,每部分任务的具体内容作为二级指标,由于在文本预处理过程中已经去除无实际意义词语及语句,因此待分析的政策文本语句都具有实际意义,并具有可比性,本文提取相似度最高的语句作为待分析样本,将语句中最大相似度数值填充到框架中来。

3.5 政策比较研究

计算结果在一定程度上可以反映出《纲要》中三大任务与各地区大数据政策文本相似性和差异性。再结合各地区实际情况,对大数据政策进行综合比较研

张涛，马海群，易扬. 文本相似度视角下我国大数据政策比较研究[J]. 图书情报工作,2020,64(12):26－37.

究,最终提出大数据发展科学化的政策建议。

4 实证研究

对我国22个地区大数据政策发布时间分析可知(见表1),重庆市和贵州省大数据政策制定时间较早,分别在2013年和2014年出台了《重庆市大数据行动计划》和《贵州省大数据产业发展应用规划纲要(2014－2020年)》,也正是因为这些地区大数据政策的先行,推动了《纲要》的出台。在《纲要》颁布后,贵州省在2016年1月通过了《贵州省大数据发展应用促进条例》,这是《纲要》颁布后我国首部大数据地方性法规,该条例不仅体现出贵州省奋力开创大数据产业发展新局面的决心,还对各地区大数据政策出台起到了重要的推进作用[19]。在随后2016年到2019年,各地区根据《纲要》中三大任务结合本地区的实际情况陆续出台了大数据实施方案或行动计划,这些政策是各地区推进大数据产业发展的引领性文件,在对大数据产业发展研究上具有代表性。

表1 各地区省级政府大数据政策文本(按发表时间排序)

年份	地区	政策名称
2013	重庆	重庆市大数据行动计划
2014	贵州	贵州省大数据产业发展应用规划纲要(2014－2020年)
2016	贵州	贵州省大数据发展应用促进条例
2016	北京	北京市大数据和云计算发展行动计划(2016－2020年)
2016	上海	上海市大数据发展实施意见
2016	广东	广东省促进大数据发展行动计划(2016－2020年)
2016	广西	广西壮族自治区促进大数据发展行动方案
2016	山东	山东省人民政府关于促进大数据发展的意见
2016	浙江	浙江省促进大数据发展实施计划
2016	江苏	江苏省大数据发展行动计划
2016	湖北	湖北省大数据发展行动计划(2016－2020年)
2016	福建	福建省促进大数据发展实施方案(2016－2020年)
2016	海南	海南省促进大数据发展实施方案
2017	山西	山西省大数据发展规划(2017－2020年)
2017	云南	关于重点行业和领域大数据开放开发工作的指导意见
2017	江西	江西省大数据发展行动计划
2017	内蒙古	内蒙古自治区大数据发展总体规划(2017－2020年)
2018	河南	河南省大数据产业发展三年行动计划(2018－2020年)
2018	河北	河北省大数据产业创新发展三年行动计划(2018－2020年)
2018	四川	四川省促进大数据发展工作方案
2019	天津	天津市促进大数据发展应用条例
2019	湖南	湖南省大数据产业发展三年行动计划(2019－2021年)
2019	黑龙江	“数字龙江”发展规划(2019－2025年)

4.1 基于《纲要》三大任务的政策比较分析

《纲要》是指导我国大数据产业发展的顶层设计,其中三大任务是大数据产业从理论研究走向实际应用的关键部分,把《纲要》三大任务中的具体内容作为参照,与各地区大数据政策文本作比较分析有利于挖掘大数据产业发展过程中的重点任务及区域特色[20],对三大任务中具体内容分析如下。

4.1.1 数据开放

任务1:数据开放过程中主要涉及到数据共享、数据资源开放、基础设施建设、宏观调控、政府治理、商事服务、安全保障、民生服务八项具体内容。结合表2对部分具体内容进行分析。

(1)数据资源开放。在该部分中强调要在依法加强安全保障和隐私保护的前提下,稳步推动公共数据资源开放。此部分数值范围是(0.402 8－0.922 1),其中广东地区最为突出,数值为0.922 1,广东省在2016年印发《广东省促进大数据发展行动计划(2016－2020年)》的通知中明确提出在依法加强数据安全保障和隐私保护的前提下,开展公共数据资源开放应用并制定政府数据资源开放的计划、目录和标准规范及安全保护准则,建设全省政府数据统一开放平台,统筹管理可开放的政府数据资源,提供面向公众的政府数据服务。数据分析显示:北京、贵州、江苏等地区数值较高,分别为0.865 0、0.861 1、0.832 9,这些地区的政策中数据资源开放层面提及较为明确。由于政府数据开放是推动大数据产业发展的基础,因此在各地区政策制定中都占重要位置。

(2)基础设施建设。强调要结合国家政务信息化工程建设规划,统筹政务数据资源和社会数据资源,布局国家大数据平台、数据中心等基础设施。此部分数值范围是(0.307 8－1),福建地区数值最高,为1,2016年福建省印发的《福建省促进大数据发展实施方案(2016－2020年)》通知中明确提出加快构建省市两级基础平台建设及推动国民经济动员大数据应用,对两项政策内容对比发现,该项内容有一部分完全吻合。数据分析显示:四川、广东等地区数值较高,分别为0.821 6、0.795 2,这些地区在政策制定过程中对基础设施建设也比较关注。很多地区把统筹政务数据资源和社会数据资源,布局区域大数据平台、数据中心等基础设施作为区域建设的重点。

(3)商事服务。在该部分中强调要鼓励政府部门高效采集、有效整合并充分运用政府数据和社会数据,掌握企业需求,推动行政管理流程优化再造,在注册登记、市场准入等商事服务中提供更加便捷有效、更有针对性的服务。此部分数值范围是(0.275 3－1), 江苏

表 2 数据开放相似度数值对比

年份	地区	任务 1:加快政府数据开放共享,推动资源整合,提升治理能力							
		八项具体内容							
		数据共享	数据资源开放	基础设施建设	宏观调控	政府治理	商事服务	安全保障	民生服务
2013	重庆	0.2120	0.755 3	0.637 2	0.222 5	0.226 1	0.345 8	0.219 0	0.265 3
2014	贵州	0.359 9	0.861 1	0.435 3	0.277 2	0.419 8	0.423 5	0.212 0	0.283 6
2016	北京	0.443 4	0.865 0	0.496 6	0.268 4	0.740 9	0.610 2	0.540 2	0.309 0
2016	上海	0.327 8	0.539 9	0.450 0	0.224 8	0.495 6	0.361 8	0.233 2	0.502 8
2016	广东	0.371 9	0.922 1	0.795 2	0.659 3	0.685 8	0.985 1	0.734 5	0.629 6
2016	广西	0.640 0	0.622 8	0.538 3	0.513 0	0.358 1	0.607 2	0.414 8	0.570 0
2016	山东	0.362 3	0.699 3	0.566 0	0.232 4	0.409 0	0.497 8	0.438 4	0.280 6
2016	浙江	0.419 8	0.716 3	0.573 9	0.464 3	0.456 7	0.563 1	0.654 6	0.406 5
2016	江苏	0.430 5	0.832 9	0.614 1	0.302 9	0.310 5	1.000 0	0.196 1	0.441 5
2016	湖北	0.481 5	0.761 4	0.597 1	0.635 5	0.475 3	0.533 1	0.298 6	0.554 0
2016	福建	0.316 0	0.593 0	1.000 0	0.581 3	0.703 2	0.491 7	0.281 0	0.466 6
2016	海南	0.466 1	0.736 9	0.524 6	0.488 0	0.387 6	0.650 2	0.208 6	0.323 9
2017	山西	0.594 1	0.799 5	0.553 9	0.513 0	0.485 1	0.456 6	0.343 2	0.378 2
2017	云南	0.271 6	0.674 6	0.327 4	0.195 5	0.203 5	0.275 6	0.438 4	0.289 6
2017	江西	0.376 7	0.425 9	0.370 7	0.299 7	0.638 3	0.564 3	0.303 1	0.258 0
2017	内蒙古	0.649 4	0.703 6	0.514 4	0.404 0	0.477 0	0.473 8	0.372 5	0.305 2
2018	河南	0.344 0	0.402 8	0.307 8	0.199 0	0.274 9	0.458 7	0.227 3	0.282 7
2018	河北	0.295 0	0.542 2	0.401 7	0.268 1	0.481 6	0.423 5	0.626 1	0.283 9
2018	四川	0.416 9	0.548 6	0.821 6	0.456 0	0.442 4	0.531 7	0.282 2	0.384 0
2019	天津	0.319 0	0.620 7	0.333 3	0.230 4	0.255 1	0.490 3	0.209 5	0.606 7
2019	湖南	0.459 8	0.405 1	0.371 3	0.222 5	0.320 7	0.275 3	0.285 4	0.223 6
2019	黑龙江	0.433 2	0.481 5	0.438 8	0.304 3	0.434 0	0.530 6	0.363 5	0.530 8
平均值		0.408 7	0.659 6	0.530 4	0.361 9	0.440 1	0.525 0	0.358 3	0.389 8

地区数值最高,为 1, 在 2016 年江苏省印发《江苏省大数据发展行动计划》的通知中所提出的有针对性的商事服务与《纲要》中推进商事服务便捷化完全吻合。数据分析显示:广东地区数值也较高,为 0.985 1,除了江苏省和广东省对推进商事服务便捷化提出了明确界定,其它省市提及较少,由于商事服务是依托于区域经济发展状况,这两个地区数值较高是与其经济较为发达有着密切的联系。

(4)安全保障。在任务 1 中单独提出了促进安全保障高效化,在数据开放共享的同时,提高公共安全保障能力,推动构建智能防控、综合治理的公共安全体系也需要重点关注。此部分数值范围是(0.196 1 – 0.734 5),数值相对较低,广东地区数值最高,为 0.734 5,在广东省印发的《广东省促进大数据发展行动计划(2016 – 2020 年)》的通知中提出在法律许可和确保安全的前提下,加强对社会治理相关领域数据流通、数据归集、数据发掘及关联分析,为妥善应对和有效处置重大突发公共事件提供数据支撑。数据分析发现:大多数地区均是在任务 3 中进行明确要求,而在任务 1 中多是提出总体性的安全保障要求,因此导致此部分数值较低。而广东省在政策制定层面非常重视数据安全保障,在任务 1 中提出安全保障也说明政府关注到了数据开放共享与数据安全保障间协同性的问题。

4.1.2 创新发展

任务 2:创新发展中主要涉及到工业大数据、新兴产业大数据、农业农村大数据、万众创新大数据、基础研究和核心技术攻关、大数据产品体系、大数据产业链七项具体内容。结合表 3 对部分具体内容进行分析。

(1)工业大数据。在该部分中强调要推动产业创新发展,培育新兴产业,助力经济转型的重要组成部分。此部分数值范围是(0.265 0 – 0.963 0),数值跨度较大,其中广西和四川的数值较高,分别为 0.930 4 和 0.963 0,2016 年广西壮族自治区人民政府发布了《促进大数据发展行动方案》,2018 年四川省也发布了《四川省促进大数据发展工作方案》,这两部政策文件在工业大数据应用试点和打造"互联网 + 智能制造"工业大数据应用基地方面关注较多。数据分析显示:北京和江苏数值也较高, 分别为 0.847 8 和 0.786 2, 这些

表3 创新发展及安全保障部分相似度数值对比

年份	地区	任务2:推动产业创新发展,培育新兴业态,助力经济转型							任务3:强化安全保障	
		七项具体内容							两项具体内容	
		工业大数据	新兴产业大数据	农业农村大数据	万众创新大数据	基础研究和核心技术攻关	大数据产品体系	大数据产业链	安全保障体系	安全支撑
2013	重庆	0.371 8	0.354 4	0.303 9	0.257 4	0.394 3	0.302 5	0.768 2	0.500 0	0.385 9
2014	贵州	0.388 2	0.390 2	0.411 1	0.570 8	0.383 6	0.418 5	0.647 4	0.540 8	0.275 6
2016	北京	0.847 8	0.642 7	0.649 3	0.578 5	0.273 4	0.292 1	0.324 2	0.753 6	0.676 3
2016	上海	0.625 5	0.710 6	0.394 7	0.556 9	0.405 3	0.310 2	0.449 4	0.507 6	0.429 5
2016	广东	0.460 0	0.606 8	0.435 4	0.685 7	1.000 0	0.765 4	0.718 1	1.000 0	0.851 1
2016	广西	0.930 4	0.399 8	0.685 0	0.671 0	0.389 9	0.309 4	0.533 4	0.353 9	0.535 4
2016	山东	0.725 0	0.337 3	0.499 3	0.426 4	0.389 9	0.611 3	0.450 4	0.493 4	0.424 6
2016	浙江	0.463 5	0.415 4	0.734 4	0.426 4	0.389 9	0.466 1	0.634 8	0.672 5	0.543 9
2016	江苏	0.786 2	0.528 9	0.379 9	0.604 2	0.358 9	0.466 0	0.621 5	0.409 6	0.433 0
2016	湖北	0.725 0	0.435 6	0.681 2	0.483 9	0.389 9	0.338 3	0.495 9	0.636 5	0.562 2
2016	福建	0.725 0	0.823 5	0.761 1	0.685 2	0.559 0	0.691 4	0.861 0	0.755 7	0.420 0
2016	海南	0.533 0	0.372 7	0.390 3	0.685 2	0.343 8	0.278 8	0.404 0	0.503 1	0.420 0
2017	山西	0.688 7	0.567 3	0.608 5	0.445 3	0.445 3	0.563 0	0.468 6	0.966 6	0.707 6
2017	云南	0.265 0	0.405 8	0.296 0	0.290 7	0.334 9	0.335 9	0.402 4	0.486 7	0.384 5
2017	江西	0.725 0	0.601 3	0.394 7	0.429 7	0.407 6	0.365 2	0.532 2	0.558 8	0.355 2
2017	内蒙古	0.565 7	0.491 5	0.444 4	0.500 6	0.427 3	0.498 4	0.514 3	1.000 0	0.939 3
2018	河南	0.625 5	0.555 5	0.319 9	0.465 5	0.493 2	0.427 8	0.403 3	0.402 1	0.425 2
2018	河北	0.625 5	0.429 7	0.394 7	0.426 6	0.340 4	0.446 4	0.448 5	0.564 4	0.441 4
2018	四川	0.963 0	0.600 9	0.500 4	0.541 4	0.517 2	0.327 4	0.826 6	0.755 7	0.454 6
2019	天津	0.417 5	0.409 1	0.440 7	0.357 5	0.297 5	0.258 4	0.377 8	0.631 4	0.441 5
2019	湖南	0.487 6	0.549 1	0.402 3	0.500 6	0.528 8	0.526 4	0.479 8	0.705 4	0.441 4
2019	黑龙江	0.441 9	0.487 1	0.536 9	0.484 9	0.389 9	0.387 5	0.638 4	0.627 0	0.627 4
平均值		0.608 5	0.505 2	0.484 7	0.503 4	0.430 0	0.426 7	0.545 5	0.628 4	0.508 0

地区比较重视工业大数据产业的发展。而从数据上看在其它地区政策中对工业大数据提及较少,可以侧面说明不同地区大数据产业发展定位的差异性。

(2)农业农村大数据。在该部分中强调要构建面向农业农村的综合信息服务体系,为农民生产生活提供综合、高效、便捷的信息服务,缩小城乡数字鸿沟,促进城乡发展一体化。此部分数值范围是(0.296－0.761 1),福建和浙江数值较高,分别为0.761 1和0.734 4,2016年,福建省发布了《福建省促进大数据发展实施方案(2016－2020年)》的通知,同年,浙江省也发布了《浙江省促进大数据发展实施计划》,这两部政策法规都明确提出了对加快农业农村大数据发展的计划。数据分析显示:广西和湖北数值较高,分别为0.685 0和0.681 2,这些地区对农业农村大数据产业的发展非常重视,并且从政策层面对农业农村大数据较为关注,如广西壮族自治区建设的智慧农庄信息管理平台在农业农村大数据、扶贫大数据建设上特点较为突出,并初见成效,这切实体现出区域特色大数据产业的蓬勃发展。

(3)基础研究和核心技术攻关。在该部分中强调要围绕数据科学理论体系、大数据计算系统与分析理论等重大基础研究进行前瞻布局,开展数据科学研究,引导和鼓励在大数据理论、方法及关键应用技术等方面展开探索。此部分数值范围是(0.273 4－1),广东的数值最高,为1,在广东省发布的《广东省促进大数据发展行动计划(2016－2020年)》中提出推动大数据核心技术攻关和产业化应用,重点突破大规模数据采集和预处理。数据分析显示:其它地区的数值普遍较低,在0.27－0.55之间,这些地区对基础研究和核心技术攻关提及较少,且关注度不高,这主要因为有些地区受限于经济发展状况,并不具备大数据核心技术攻关的综合实力。

(4)大数据产业链。在该部分中强调要支持企业开展基于大数据的第三方数据分析发掘服务、技术外包服务和知识流程外包服务,鼓励企业根据数据资源基础和业务特色,积极发展互联网金融和移动金融等

新业态。此部分数值范围是(0.324 2 - 0.861 0),福建、四川、重庆、广东这 4 个地区数值相对较高,分别为 0.861 0、0.826 6、0.768 2、0.718 0,这些地区在政策中不同程度都提到了完善大数据产业链措施。数据分析显示:其它地区的数值相对较低(0.32 - 0.63),对该部分内容提及较少。由于在此部分中,多数地区在政策中对相关内容提及较为宏观,且特征词分散,导致文本相似度数值较低。

4.1.3 安全保障

任务 3:安全保障中主要涉及到健全大数据安全保障体系和强化安全支撑两项具体内容,结合表 3 对具体内容进行分析。

(1)安全保障体系。在该部分中强调要加强大数据环境下的网络安全问题研究和基于大数据的网络安全技术研究,落实信息安全等级保护、风险评估等网络安全制度,建立健全大数据安全保障体系。此部分数值范围是(0.353 9 - 1),广东、内蒙古、山西等地区的数值较高,分别是 1、1、0.966 6,以上 3 个地区从政策层面凸现了对安全保障体系的重点关注。数据分析显示:虽然个别地区对健全安全保障体系提及较少,但整体来看国家及各地区还是非常关注大数据安全问题的,数据安全是国家安全的重要基础,因此各地区在发展大数据产业的同时对数据安全保障都极为重视。

(2)安全支撑。在该部分中对网络安全及防护提出了明确要求,强调要采用安全可信产品和服务,提升基础设施关键设备安全可靠水平。此部分数值范围是(0.275 6 - 0.939 3),广东、内蒙古数值较高,分别为 0.939 3 和 0.851 1,其中广东省作为大数据产业的领跑省份,对数据安全支撑方面尤为重视,而内蒙古自治区在 2017 年发布的《内蒙古自治区大数据发展总体规划(2017 - 2020 年)》中重点强调了提升大数据安全保障能力,要把网络安全作为大数据发展的重要前提,健全安全保障体系,提升技术支撑能力,切实保障数据安全。数据分析显示:贵州的数值为 0.275 6,虽然最低,但是贵州省在 2019 年出台的《贵州省大数据安全保障条例》是我国大数据安全保护省级层面的首部地方性法规,是贵州省大数据产业发展制度保障顶层设计的又一项重要成果。健全大数据安全保障体系和强化安全支撑是相辅相成的,数据安全是数据开放共享的前提条件,因此在大数据产业发展过程中,强化安全支撑是首要任务。

4.2 《纲要》三大任务间政策比较分析

通过对《纲要》中三大任务在各地区相似度平均值分析,可以探寻大数据政策重点内容及各地区政策的共性,按照数值将其划分为三档,如图 3 所示:

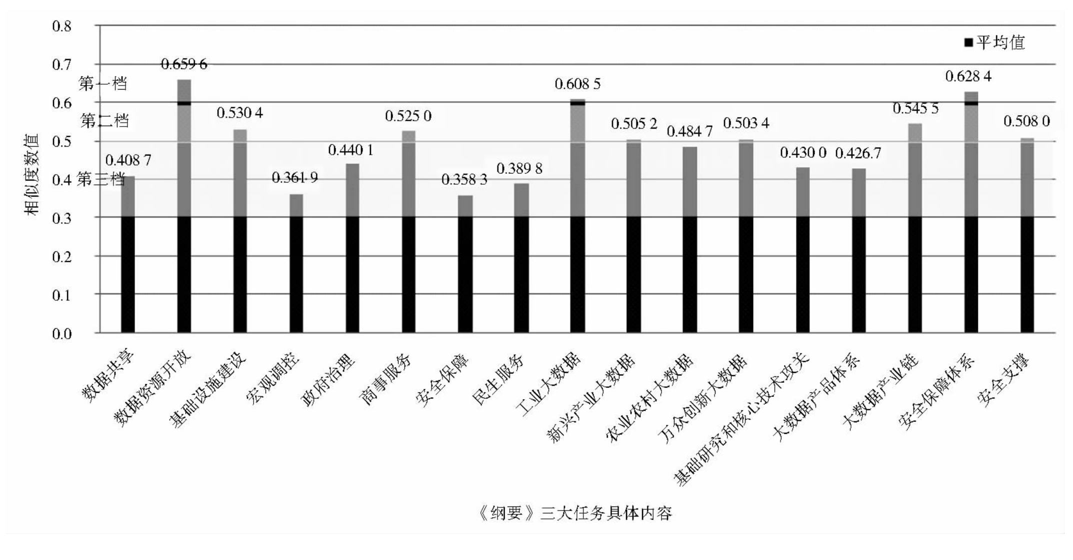


图 3 《纲要》三大任务中具体内容相似度平均值

第一档:数值在 0.6 - 0.7 之间,包括数据资源开放 0.659 6、安全保障体系 0.628 4、工业大数据0.608 5。数据分析显示:这些内容在各地区政策中关注度最高,在大数据政策制定过程中,数据资源开放共享和安全保障体系的构建长期以来都是政策制定所需要重点关注的对象,工业是国家经济发展的基础,工业大数据创新发展是实现智能制造的重要抓手,因此从政策制定层面多数地区比较重视工业大数据的发展。

第二档:数值在 0.5 - 0.6 之间,包括基础设施建设 0.530 4、商事服务 0.525 0、新兴产业大数据

0.505 2、万众创新大数据 0.503 4、大数据产业链 0.545 5、安全支撑 0.508 0。数据分析显示：《纲要》中的部分内容相对较宏观且有些任务受地域影响较大，各地区会根据区域特点来制定大数据战略，因此数值波动较大，如在经济欠发达地区，基础设施建设较为落后，新兴产业大数据的发展则相对迟缓。

第三档：数值在 0.3－0.5 之间，包括数据共享 0.408 7、宏观调控 0.361 9、政府治理 0.440 1、安全保障 0.358 3、民生服务 0.389 8、农村大数据 0.484 7、基础研究和核心技术攻关 0.430 0、大数据产品体系 0.426 7。数据分析显示：安全保障、宏观调控、民生服务数值最低，很多地区并没有在数据开放共享部分体现数据安全保障，这并非不重视安全问题，大多数地区是在强化安全保障部分中对数据安全做了明确要求。而宏观调控、民生服务等任务则体现了《纲要》中政策引领性、全面性特点，虽然在省级大数据政策中对这些方面提及较少，但很多地区都单独出台了与民生服务、政府治理等相关的专项大数据政策。

4.3 地区间政策比较分析

4.3.1 各地区政策平均相似度比较分析

《纲要》三大任务与各地区大数据政策平均相似度较高可以体现出大数据政策制定的完整性与全面性，将此部分划分为三档，如图 4 所示：

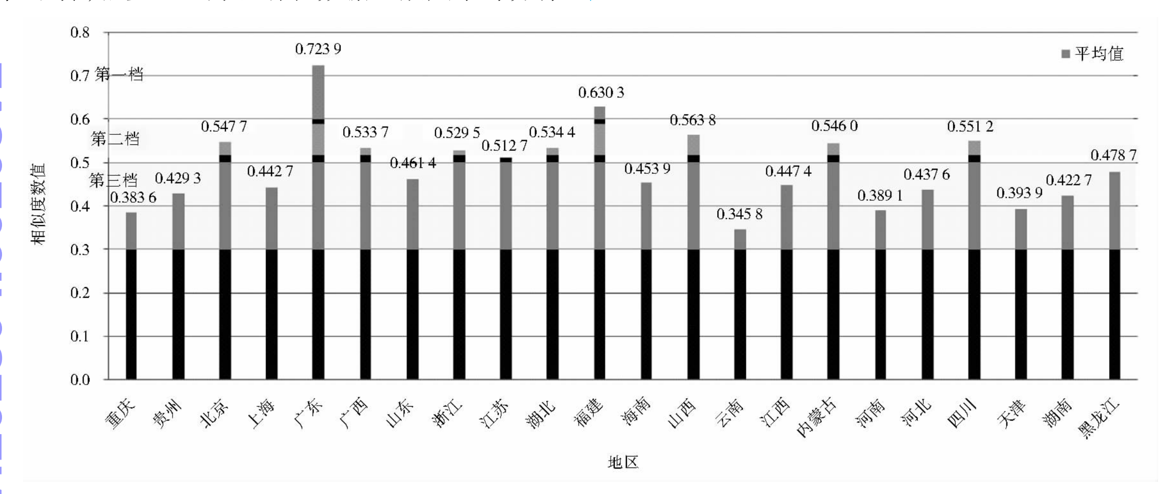


图 4　各地区大数据政策平均相似度数值

（1）第一档：数值在 0.6－0.8 之间，以广东、福建最为突出，这两个地区主要特点是处于沿海地区，GDP 总值较高，经济较为发达。广东省在 2016 年发布的《促进大数据发展行动计划（2016－2020 年）的通知》与《纲要》中文本相似度比较数值最高为 0.723 9。广东省是国内率先关注并推动大数据的地区之一，作为工业、制造业强省，大数据领域企业聚集地，广东省具有发展大数据产业独特的优势。福建省在 2016 年发布了《福建省促进大数据发展实施方案（2016－2020 年）》的通知，数值为 0.630 3，仅次于广东省，福建省通过加速"数字福建"的建设来抢占数字经济的前沿，依托高校建立的大数据基础技术研究基地及大数据研究院，为区域大数据产业发展提供了强有力的支撑。

（2）第二档：数值在 0.5－0.6 之间，主要有北京、江苏、浙江、湖北、四川、山西、广西、内蒙古等地区，将这些地区分为两种情况：①根据连玉明[21]《中国大数据发展报告》中的大数据发展总指数看，北京、江苏、浙江等地区大数据产业发展较好，但从数值上并未体现出来。②部分地区如广西、内蒙古、四川等根据区域特点来制定大数据政策。广西壮族自治区通过智慧农庄信息管理平台，使该地区在农业农村大数据建设层面初见成效。内蒙古自治区通过打造云计算和大数据产业集群，逐步建设成为中国北方大数据中心。四川省德阳市通过发展工业大数据应用与服务，推动建立智能制造集群，形成"互联网＋智能制造"工业大数据应用基地。

（3）第三档：数值在 0.3－0.5 之间，主要有上海、天津、重庆、山东、河北、河南、湖南、云南、贵州等地区，根据连玉明[21]《中国大数据发展报告》和自建语料库中的政策分析将这些地区分为两种情况：①大数据产业发展较好的地区，这些地区基于《纲要》发布了很多大数据专项政策，有些内容在各省级政府大数据政策中并没有全部体现，因此从数据层面表现不足，如上海、贵州、重庆、山东等地区。②大数据产业发展相对较缓慢的地区，很多大数据相关基础设施尚无法满足，因此无法从数据层面体现，这主要集中在经济欠发达

地区。

进一步对第三档中部分数值较低的地区做如下分析:

贵州地区平均值为 0.429 3,但综合分析贵州省大数据产业发展情况可知:贵州省 2013 年就走上了大数据之路,如今已经成为大数据时代的领跑者,根据团队自建语料库数据统计,贵州省自 2014 年起发布省级政府层面的大数据政策 10 部,各地市共发布大数据相关政策 70 多部,是全国大数据政策内容制定最细致、最完善、最丰富的地区。本文选取的政策是《贵州省大数据产业发展应用规划纲要(2014 - 2020 年)》,虽然从相似度数值上看较低,但贵州省采取政策群的方式来推动大数据产业在该地区的发展,并收到了较好的效果。如《贵州省发展农业大数据助推脱贫攻坚 3 年行动方案(2017 - 2019 年)》《贵州省人民政府办公厅关于深入推进政务服务领域大数据和人工智能集成应用的实施意见》,这些都是单独针对《纲要》中具体任务进行专项部署的政策文件。

重庆地区平均值为 0.383 6,由于重庆市是全国制定大数据政策最早的地区,要早于《纲要》的发布,因此该政策与《纲要》对比数值较低是可以理解的,但这不代表重庆市大数据产业发展速度缓慢。重庆市的大数据智能化产业已经初具规模,建设的大数据产业园、仙桃数据谷已形成了具有国际竞争力的创新生态圈,重庆市在政府管理、智能交通、智能物流等领域的大数据智能化应用水平全国领先。

云南地区平均值为 0.345 8,虽然最低,但任务 1 中数字资源开放的数值较高为 0.674 6。云南省是一个集边疆、山区、贫困等不利因素为一体的欠发达省份,因此在大数据发展过程中更多关注数字资源开放工作,2017 年云南省人民政府发布的《关于重点行业和领域大数据开放开发工作的指导意见》是省级层面的大数据政策,该政策内容主要以云南省重点行业和领域大数据的开放开发工作为主。

4.3.2 《纲要》三大任务与各地政策相似度比较分析

通过《纲要》三大任务与各地区政策相似度分项比较有助于发现在地区间政策制定的差异,进而挖掘大数据产业发展的区域特色,数据分析发现任务 3 安全保障是各地区政策制定中最为关注的内容,具体分析如下:

(1)任务 1:数据开放中各地区平均数值为 0.459 2,数值最低,如图 5 所示,广东省此部分数值较为最高,为 0.722 9,广东省明确提出建设全省统一电子政务数据中心和政务信息资源共享平台的目标。其它地区在不同程度上提及到数据开放共享,但是整体数值较低的原因主要是由于在宏观调控、政府治理等方面平均数值较低,其它地区中对该部分内容的描述都没有广东省颁布的《广东省促进大数据发展行动计划(2016 - 2020 年)的通知》中的内容细致,因此影响了整体数值。

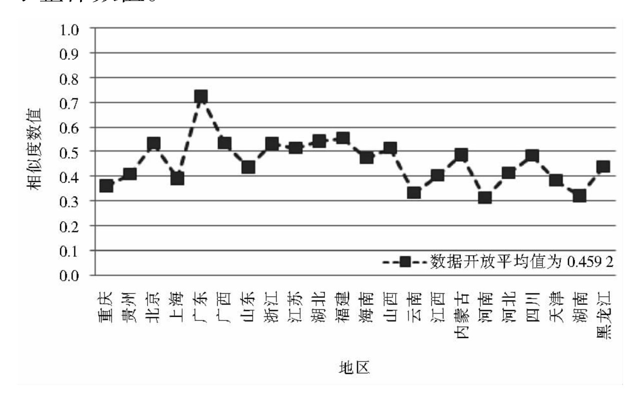


图 5 任务 1 数据开放与各地区政策相似度数值

(2)任务 2:创新发展中各地区平均数值为0.500 6,如图 6 所示,以福建(0.729 5)、广东(0.667 3)最为突出。由于国家鼓励建设有区域特色的大数据产业,要发挥地区特色,因此各地区关注点有所差别,这成为平均值略低的因素之一,而构建工业大数据、新兴产业大数据、农业农村大数据都需要大数据底层建设基础,福建省、广东省大数据基础设施建设较好,并且建设资金较充裕,在创新发展方面可以投入更多精力来推动大数据产业在新兴行业的发展。

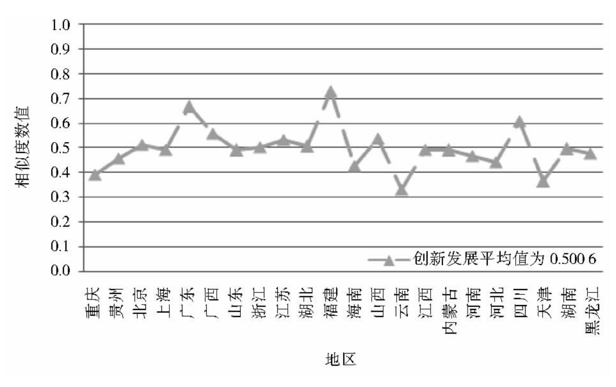


图 6 任务 2 创新发展与各地区政策相似度数值

(3)任务 3:安全保障中各地区平均数值最高为 0.568 2,如图 7 所示,以内蒙古(0.969 7)、广东(0.925 6)、山西(0.837 1)最为突出,由此可见各地区在政策制定过程中极为重视数据安全问题,以内蒙古自治区为例,作为国家大数据综合试验区,近年来内蒙古自治区对大数据产业发展尤为重视,2017 年发布了

《内蒙古自治区大数据发展总体规划(2017－2020年)》，其中数据安全保障体系与《纲要》中安全保障部分相似度高，该地区现存大数据安全平台建设滞后、数据开放共享程度不高、相关政策法规不完善等问题，因此从政策层面可以体现出内蒙古自治区对安全保障重视程度较高，也能看出内蒙古自治区建立世界级大数据中心的决心。

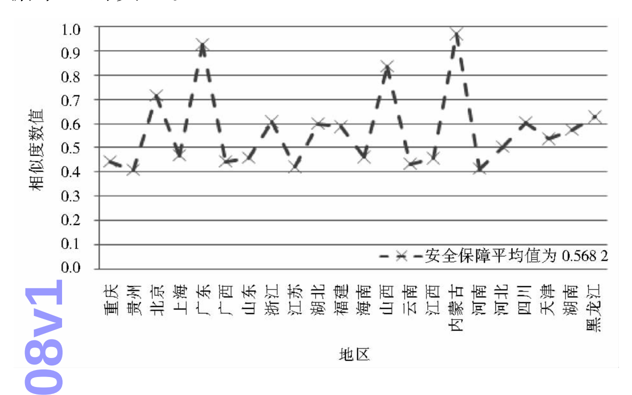


图 7　任务 3 安全保障与各地区政策相似度数值

5　结论与建议

本文以文本相似度为视角对《纲要》中三大任务与各地区大数据政策进行了综合比较研究，广东省、福建省从数值上看总体表现最好，从整体情况来看，由于地方政府重视程度较高，大数据产业呈现出发展迅速、重点任务明确、区域特色突出等特点。不同地区的大数据政策中对数据资源的开放和安全保障体系关注度均较高，这体现出各地区大数据政策制定的相似性。图 5 中内蒙古自治区、四川省等地区在安全保障与创新发展中数值较高，这体现出不同地区大数据政策制定的差异性。在所收集的政策文本中并没有辽宁省、新疆维吾尔自治区、西藏自治区等地区的省级政府层面的大数据文件。结合自建语料库所收集的语料发现，辽宁省各地市大数据政策文件共 104 条，数量上在全国也名列前茅，沈阳市、大连市制定的政策较多，结合数据分析可知，辽宁省的地市大数据规划先行，而省级政府层面大数据规划却相对滞后。基于以上结论，笔者对我国大数据政策提出建议：

5.1　提升地方经济水平可以促进大数据产业快速发展

从综合数据来看，广东省和福建省平均值最高，以广东省最为突出，数值为 0.723 9。结合自建语料库所收集的政策文本分析发现，近年来广东省共发布数据类政策 117 条，在政策发布数量上领跑全国，广东省作为我国的经济强省，作为国家改革开放的窗口，在中国区域经济发展全面进入城市群引领时代后，拥有广深两座一线城市，在聚集人才、资金、产业的能力上，要领先于其它地区，这也是广东省能够全面快速推动大数据产业创新发展的重要因素之一[22]。福建省通过出台一系列大数据政策来推动实体经济与数字经济携手并进，推动传统产业智能升级，还依托厦门大学、福州大学建立大数据基础技术研究院，这些都为福建省的大数据产业发展提供了有力的支撑。由于大数据基础设施建设、基础研究及核心技术研究会受到地方经济直接影响，而它又是大数据产业发展基础，因此地方经济是影响大数据产业发展的重要因素。

5.2　大数据开放共享体系与安全保障体系建设是最重要的基础工程

数据开放与安全保障作为《纲要》中最重要的两大任务，它们在各地区的落实与执行情况至关重要。数据资源开放数值范围是(0.402 8－0.922 1)，平均数值为 0.659 6，安全保障体系数值范围是(0.353 9－1)，平均数值为 0.628 4，这说明不同地区都把数据资源开放和安全保障体系作为最关注的内容，也体现了各地区大数据政策制定的相似性。开放共享是大数据的核心价值，数据共享开放的程度是国家数字经济竞争力的决定要素，要加快建立统一的大数据开放共享标准体系，整合大数据资源的数据标准和应用规则[23]。大数据开放共享的同时要加强数据安全防护意识，在各地区大数据政策中对数据安全关注度较高，但目前尚缺乏完整的政策体系来保障大数据安全。因此大数据开放共享标准体系建设和数据安全保障体系建设是国家及地方政府最为重视的两项基础工程。

5.3　发挥地域优势构建有特色的大数据产业

我国各地区大数据产业发展不均衡，受地域影响较大，如内蒙古自治区政策整体平均值为 0.545 9，相对较低，但任务 3 安全保障的数值高达 0.969 7，最为突出，其原因是内蒙古自治区全力建设我国北方地区的云计算和大数据中心，因此对于数据的安全保障尤为重视，从健全大数据安全保障体系和提升大数据安全技术支撑能力两个方面全力保障云计算和大数据中心的数据安全。四川省数值为 0.551 2，但工业大数据部分数值为 0.963 0，表现最为突出，四川省在政策制定中突出特色，通过区域特色来推进大数据产业的实施，以德阳市为例，该市通过工业大数据应用与服务，推动建立智能制造集群，形成“互联网＋智能制造”工业大数据应用基地。贵州省整体平均值仅为 0.429 3，虽然数值较低，但贵州省所发布大数据相关

配套政策较多,内容完善且覆盖面广,已初具规模,并形成区域大数据政策群,综合分析贵州省大数据产业发展特色鲜明,有引领示范作用[24]。构建有区域特色的大数据产业体现了各地区大数据政策制定的差异性,因此各地区应发挥优势,大力发展有区域特色的大数据产业。

5.4 加快人工智能与实体经济融合来推动大数据产业发展

2017年7月国务院发布《新一代人工智能发展规划》后,各地区政府先后出台了一系列人工智能相关政策文件。2019年3月,国务院总理李克强在政府工作报告中提出“智能+”的概念,与此同时还强调要深化大数据、人工智能等研发应用。“智能+”将正式接棒“互联网+”,这也意味着我国人工智能即将开启和互联网一样的规模化发展之路,在未来几年内将快速在各行业落地。2020年2月习近平在中央全面深化改革委员会第十二次会议中强调要鼓励运用大数据、人工智能等技术,在疫情监测分析、防控救治、资源调配等方面发挥支撑作用。很多地区大数据政策中已不同程度提出人工智能、智慧城市、智慧社会、智能防控等概念,如广东省、重庆市、贵州省等。在大数据产业应用进程中,各地区应准确把握全球人工智能发展态势,构建基于5G、大数据、超级计算、传感网等新理论新技术的新一代人工智能创新体系,加强人工智能应用技术研发,大力推动人工智能与实体经济深度融合,培育高端高效的智能经济,最终建设安全便捷的智能社会[25]。这些都为进一步明确未来人工智能视域下大数据产业发展方向奠定坚实基础。

单从数据分析结果上看,部分地区的数据并不能反映出大数据产业的实际发展水平,这就要结合实际情况进行综合分析。笔者将在该领域持续研究,继续收集大数据政策相关解读、报道、评论等文本,尝试用机器学习、深度学习等方法来优化模型,并在未来有针对性的对大数据政策双向评价及人工智能政策和大数据政策协同方面做前瞻性研究。

参考文献:

[1] 裴雷,孙建军,周兆韬.政策文本计算——一种新的政策文本解读方式[J].图书与情报,2016 (6):47-55.
[2] 张勇进,王璟璇.主要发达国家大数据政策比较研究[J].中国行政管理,2014(12):113-117.
[3] 汤志伟,龚泽鹏,郭雨晖.基于二维分析框架的中美开放政府数据政策比较研究[J].中国行政管理,2017(7):41-48.
[4] 王本刚,马海群.开放政府数据的政策比较研究[J].情报资料工作,2017(6):33-40.
[5] 赵远.内蒙古与十省(市)大数据政策比较研究——基于“目标-工具”二维分析框架[D].呼和浩特:内蒙古大学,2019.
[6] ZUIDERWIJK A, JANSSEN M. Open data policies, their implementation and impact: a framework for comparison[J]. Government information quarterly,2014,31(1):17-29.
[7] CHATZINIKOLAOU E, FAULWETTER S, MAVRAKI D, et al. Datapolicy and data sharing agreement in the Life Watch Greeceresearch infrastructure [J]. Biodiversity data journal, 2016 (4): e10849.
[8] TATIANA-CAMELIA D. The comparative method for policy studies: the thorny aspects[J]. Holistica -journal of business and public administration, 2019,10(1):56-67.
[9] LIN D. An information-theoretic definition of similarity [C]//Proceedings of the 15th international conference on machine learning. San Francisco: Margan Kaufmann,1998:296-304.
[10] 陈二静,姜恩波.文本相似度计算方法研究综述[J]. 数据分析与知识发现, 2017(6):1-11.
[11] 黄文彬,车尚锟.计算文本相似度的方法体系与应用分析[J].情报理论与实践, 2019,42(11):128-134.
[12] 李琳,李辉.一种基于概念向量空间的文本相似度计算方法[J]. 数据分析与知识发现, 2018(5):47-58.
[13] 曹祺,赵伟,张英杰,等.基于Doc2Vec的专利文件相似度检测方法的对比研究[J].图书情报工作, 2018,62(13):74-81.
[14] 张文萍,黎春兰. 基于文本空间表示模型的文本相似度计算研究[J].现代情报, 2013,33(2):21-24.
[15] 马海群,张涛. 文献信息视阈下面向智慧服务的语料库构建研究[J].情报理论与实践, 2019,42(6):124-130.
[16] 中国科学院计算技术研究所. ICTCLAS2016[EB/OL].[2019-09-28]. http://ictclas.nlpir.org/.
[17] 张涛,马海群. 一种基于LDA主题模型的政策文本聚类方法研究[J].数据分析与知识发现, 2018(9):59-65.
[18] 武永亮,赵书良,李长镜,等.基于TF_IDF和余弦相似度的文本分类方法[J].中文信息学报,2017,31(5):138-145.
[19] 李樵. 我国促进大数据发展政策工具选择体系结构及其优化策略研究[J].图书情报工作,2018,62(11):5-15.
[20] 张涛,蔡庆平,马海群.一种基于政策文本计算的政策内容分析方法实证研究[J]. 信息资源管理学报,2019(1):66-76.
[21] 连玉明. 中国大数据发展报告[M]. 北京:社会科学文献出版社,2019.
[22] 刘亚亚,曲婉,冯海红.中国大数据政策体系演化研究[J].科研管理, 2019,40(5):13-23.
[23] 黄如花, 温芳芳.我国政府数据开放共享的政策框架与内容:国家层面政策文本的内容分析[J]. 图书情报工作, 2017,61(20):12-25.
[24] 张会平,郭宁,汤玺楷.推进逻辑与未来进路:我国政务大数据政策的文本分析[J].情报杂志,2018(3):152-157,192.
[25] 汤志伟,雷鸿竹,周维.中美人工智能产业政策的比较研究——基于目标、工具与执行的内容分析[J].情报杂志,2019(10):73-80.

张涛，马海群，易扬. 文本相似度视角下我国大数据政策比较研究[J]. 图书情报工作,2020,64(12):26－37.

作者贡献说明：
张涛：负责论文数据的采集、主体内容撰写、实验开展和结果分析；
马海群：负责论文思路及框架构建，并在论文撰写过程中提出修改意见；
易扬：负责研究方法、实验开展与研究结果的完善与分析。

Comparative Analysis of China's Big Data Policies from the Perspective of Text Similarity

Zhang Tao[1] Ma Haiqun[2] Yi Yang[3]
[1] Information and Network Center, Heilongjiang University, Harbin 150080
[2] Research Center of Information Resource Management, Heilongjiang University, Harbin 150080
[3] Department of Mathematics, Heilongjiang University, Harbin 150080

Abstract: [**Purpose/significance**] The formulation and implementation of big data policies is an important means for the country to promote the development of the big data industry. Therefore, research on big data policies has received widespread attention from the society. [**Method/process**] From the perspective of text similarity, the article compares the Big Data Development Action Plan issued by the State Council and the texts of big data policies released in 22 regions. [**Result/conclusion**] Data shows: the policies formulated by Guangdong Province and Fujian Province are the most complete and comprehensive; open data sharing and security guarantees the highest overall attention in the formulation of big data policies in various regions, showing similarity; regional characteristics are more prominent, showing differences. With the successive release of artificial intelligence policies in various regions, future research on big data policies under the vision of artificial intelligence will become a new direction.

Keywords: text similarity big data policy policy comparative study policy text computing

《图书情报工作》2020 年选题指南

[**编者按**]本选题指南是根据本刊的定位、性质与发展需要，结合图情档学科前沿热点及当前与未来需要解决的重要问题，邀请本刊编委和青年编委为本刊策划定制，再经编辑部整理、修改和补充而形成的。这是本刊2020年度关注、报道的重点领域(包括但不限于这些选题)，供作者选题和研究以及向本刊投稿时的参考和借鉴。

1. 中国特色图情档学科体系、学术体系、话语体系建设
2. 图情档一级学科建设与融合发展战略
3. 图书馆“十四五”规划编制的关键问题
4. 国家文献信息资源保障能力及其建设
5. 开放科学背景下信息资源建设问题
6. 全民阅读中图书馆的定位与担当
7. 图书馆空间服务的理论与实践
8. 嵌入式学科服务的绩效评价与管理
9. 公众科学、科学素养与泛信息素养
10. 图书馆服务本科教育的模式与能力
11. 图书馆文化传承与文化育人的理论与实践
12. 图书馆出版与出版服务
13. 新媒体时代图书馆科学传播的功能与实践
14. 图书馆营销推广的战略与策略研究
15. 图书馆泛合作研究的实践与理论
16. 国家区域发展战略下图书馆联盟建设与创新服务
17. 网络空间治理的情报学问题
18. 知识产权信息服务能力与效果评估
19. 信息分析中的新技术与新方法
20. 情报服务标准化与评价
21. 数字人文与数字学术的研究与实践
22. 人工智能在图情档中的应用
23. 图书馆智能服务与智慧服务
24. 开放数据生态中的元数据发展模式研究
25. 开放科学数据行为及其模型构建
26. 数据资源建设与数据馆员能力建设
27. 大数据时代信息组织与知识组织
28. 科学数据管理与服务
29. 学术成果监测与学科竞争力分析
30. 情报计算(计算情报)的理论与方法
31. 情报分析服务质量与效能评价
32. 情报研究与智库研究的关系
33. 科学与技术前沿分析理论与方法
34. 健康中国2030战略下的健康信息学
35. 人机交互行为及服务模式创新
36. 图情档在新型智库建设中的作用机制
37. 智能信息服务的理论和方法
38. 数字公共文化资源、服务与体系建设
39. 数据时代政务信息资源管理和开发利用
40. 数字档案馆生态系统治理策略
41. 档案数据治理理论与治理体系
42. 政府数据开放平台应用与评价
43. 社会记忆视角下档案信息资源整理、保护与开发
44. 民族文献遗产产业化开发与利用
45. 图情档学科教育模式与人才培养能力